Received: August 21, 2023. Revised: October 4, 2023. 

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# Uninhibited Positional and Contextual Attention in Spectral-Based (SPT) Transformer with Multi-head Shortcut for Improved Remaining Useful Life Forecasting in Industry 4.0

**Abhishek Dwivedi[1]*** **Nikhat Raza Khan[2]**

[1]*Department of Computer Science & Engineering, IES University, Bhopal, India*
[2]*Department of Computer Science & Engineering, IES College of Technology, Bhopal, India*
* Corresponding author's Email: psit.abhishek@gmail.com

**Abstract:** This research paper aims to forecast equipment's remaining useful life (RUL) to improve maintenance planning and reduce costs. This paper presents the spectral-based transformer (SPT) model, designed for predicting the remaining useful life (RUL) in the evolving maintenance landscape of industry 4.0. Proactive maintenance is becoming increasingly important as it improves performance and reduces losses. SPT utilizes advanced attention mechanisms and innovations, which have been evaluated on the C-MAPSS dataset to simulate various operations. The contributions include discrete cosine transform attention (DCTA), uninhibited positional and contextual attention (UPCA), multi-head shortcuts, and bidirectional structures. Component efficacy is rigorously assessed through ablations. The results demonstrate that SPT exhibits superior performance compared to other methods, with a notable advantage on the challenging FD002 and FD004 sub-datasets within the C-MAPSS dataset. The proposed method decreases the root mean square error (RMSE) by 14% and enhances the performance scores of FD002 and FD004 by 10% and 24%, respectively. Additionally, it reduces the RMSE of FD004 by 15%. The model outperforms current methods, showing stability and generalization across different subsets of data. SPT demonstrates proficiency in capturing degradation patterns, which shows the potential for accurate remaining useful life (RUL) prediction. This tool is designed for time-series regression and has potential applications in various industries. Future research could focus on expanding the system's capabilities to process higher frequencies and broader contexts effectively. The SPT method provides a thorough approach for predicting the remaining useful life (RUL). It can potentially improve maintenance decisions and system performance in the context of Industry 4.0.

**Keywords:** Transformer, Remaining useful life, Multi-head shortcut, Spectral and temporal attention.

## 1. Introduction

The maintenance decision-making process has been transformed in the context of Industry 4.0 [1]. Maintenance planning, equipment failure threshold assessment, and machinery inspection have gained significance in industrial settings. The prompt deactivation of machines upon failure is now recognized as a preventive measure to reduce further damage. The modifications demonstrate the changing nature of Industry 4.0, highlighting the growing importance of proactive maintenance strategies and equipment monitoring. The integration of maintenance applications in factory environments is essential for fostering innovation. Monitoring machine conditions is essential for optimizing operations, predicting failures, and estimating remaining useful life (RUL) for maintenance planning.

Maintenance strategies can be categorized as proactive or reactive. Proactive maintenance aims to optimize system performance and minimize financial losses, while reactive maintenance addresses issues after failure. Maintenance strategies have transitioned from diagnostic-based error correction to a more comprehensive approach called prognostics [2]. Prognostics refers to predicting the remaining useful life (RUL) of systems or subsystems, primarily regarding time. The main objective of this initiative

is to set operational thresholds and anticipate maintenance needs in advance [3]. Estimating Remaining Useful Life (RUL) is essential for timely decision-making, cost reduction, and improved on-site maintenance effectiveness. This improves the system's overall health and safety. Remaining useful life (RUL) estimation is a valuable technique in prognosis and health management (PHM). It has gained recognition for its effectiveness in enhancing performance and minimizing financial losses. The term "RUL" denotes the operational lifespan of an asset until it becomes inoperable. This concept offers planning, cost efficiency, and operational performance advantages.

RUL prediction methods can be categorized into three groups: data-driven, physics model-based, and hybrid methods that integrate both approaches [3]. Data-driven methods leverage historical system data and fault records to extract insights and detect patterns. Physics model-based methods leverage expert knowledge and may not rely on historical data. Hybrid strategies combine elements from both approaches. This study uses data analysis to estimate turbofan jet engines' remaining useful life (RUL). The analysis employs NASA's C-MAPSS simulator-generated dataset for turbofan engine degradation simulation. SVM and similarity-based methods are appropriate for scenarios with limited or extensive historical failure data [5]. Predictive maintenance encompasses two main tasks: classification and regression. Classification algorithms categorize system states, while regression algorithms predict remaining useful life (RUL) values.

In industry 4.0 environments, accurate and timely forecasting of machinery and equipment's remaining useful life (RUL) is a critical challenge. Existing predictive maintenance methods often struggle to fully harness the potential of advanced machine learning techniques. The research problem is to improve the precision and reliability of RUL forecasting in Industry 4.0 by exploring the integration of uninhibited positional and contextual attention mechanisms within the spectral-based transformer (SPT) model, coupled with multi-head shortcut connections. Integrating maintenance decision-making, remaining useful life (RUL) estimation, and predictive maintenance strategies in the Industry 4.0 framework aims to enhance system performance, minimize operational disruptions, and mitigate economic consequences.

Our contribution are as follows:

- *Spectral attention modules based on discrete cosine transform attention (DCTA)*: We propose using DCTA instead of traditional attention methods that rely on weights and bias. DCTA utilizes the computationally efficient discrete cosine transform (DCT) to compute attention, resulting in a real frequency spectrum using fourier transform (FFT).
- *Uninhibited positional and contextual attention (UPCA):* We introduce UPCA to separate positional and contextual embeddings, reducing noise from positional embedding and enhancing attention performance.
- *Multi-head shortcut (MHS) mechanism*: We introduce a multi-head shortcut mechanism to improve feature representation and prevent feature collapse, ultimately enhancing the overall effectiveness of the model.
- *Bidirectional structure*: To enhance feature extraction along the temporal dimension, we incorporate a bidirectional structure into the model, leading to more accurate and effective RUL predictions.

The subsequent sections of this paper are organized as follows: Section 2 reviews the literature on RUL prediction. Section 3 proposes Spectral-based (SPT) Transformer model using multi-head shortcut and UPCA mechanism for RUL prediction. Section 4 describes the experimental setups, technical details, and model evaluation metrics. Section 5 presents the results of the proposed approaches and compare with state-of-art techniques. Finally, section 6 conclusion of the study.

## 2. Related works

Numerous studies have examined the application of machine learning algorithms to estimate the remaining useful life (RUL) and enable predictive maintenance for turbofan jet engines. Mathew et al. [7] found that the random forest algorithm outperformed others in estimating the remaining useful life (RUL). Ahsan et al. [8] developed an autoregressive model based on NASA's turbofan engine dataset for classification and regression tasks. Mosallam et al. employed classification and regression techniques to calculate the Remaining Useful Life (RUL) [9].

Ensemble learning is a vital technique for enhancing performance. Soni et al. [3] used LSTM and CNN deep learning techniques to forecast gas turbines' remaining useful life (RUL). The CNN approach demonstrated superior reliability in comparison to LSTM. Al-Dulaimi et al. [10] utilized LSTM and CNN algorithms to improve the performance of their hybrid neural network model for predicting remaining useful life (RUL). Ellefsen et al.

proposed a semi-supervised model to investigate the impact of limited labeled training data on estimating remaining useful life (RUL). Several studies have recognized the superior performance of the convolutional neural network (CNN) approach compared to the long short-term memory (LSTM) method [12-14].

Attention mechanisms have been suggested as an alternative to conventional CNN architectures in LSTM-based research on predictive maintenance. Citations [15] and [16] are referenced. CNNs excel at capturing local patterns in time-series data. However, they may face challenges in managing long-term dependencies and accurately identifying crucial segments for precise predictions. Attention is a mechanism that assigns importance to segments of an input sequence based on their relevance to predictions. Recent studies have shown that attention-based LSTM models improve performance in industrial settings for fault detection and failure prediction [9, 17].

Zhang et al. presents the dual aspect self-attention based on the transformer (DAST) method, specifically developed for efficient remaining useful life (RUL) prediction. DAST employs dual parallel encoders to extract sensor and time step information through self-attention without using RNN/CNN components [20]. Wang et al. introduces a novel joint deep learning architecture that combines a transformer encoder with a temporal convolution neural network (TCNN). The transformer model can capture long-range dependencies, while the TCNN model focuses on extracting local features. The two parts are trained in a regression module, which sets them apart from traditional ensembles. The model performs exceptionally well on the C-MAPSS [19] dataset, particularly in challenging scenarios. Limitations can occur in more straightforward situations due to transformer overfitting [21]. Xu et al. presents an improved Transformer-based method for predicting remaining useful life (RUL). The method is designed to handle diverse operating conditions and high-dimensional sensor data. This approach incorporates attention mechanisms and deep learning to consider spatio-temporal characteristics and conditions. Data preprocessing techniques such as clustering and standardization eliminate variations caused by different conditions. Self-attention effectively captures spatio-temporal features while preserving the integrity of the original sequences [22]. Zhou et al. highlights the importance of integrating domain knowledge about equipment degradation into machine learning models to improve the accuracy of remaining useful life (RUL) predictions. The approach consists of three steps. The study involves the identification of knowledge sources, formalization using Piecewise and Weibull expressions, and integration into the machine learning pipeline [23].

## 3. Proposed methodology

The spectral-based transformer (SPT) uses Transformer and BiLSTM architectures to estimate Remaining Useful Life (RUL) efficiently. To capture complicated temporal correlations in time series data while resolving transformer constraints.

Our transformer variation, uninhibited positional and contextual attention (UPCA) captures temporal attention via a 2D-tensor representation. Contextual embedding and residual connections preserve input sequence temporal information. Due to time series data noise sensitivity, we avoid absolute pooling encoding (APE) [18], unlike the transformer. Contextual embedding uses discrete cosine transformation and multi-head shortcuts. Concatenation, addition, and normalization before a fully connected feed-forward layer integrates output with temporal attention. This layer employs the exponential linear unit (ELU) activation function for smoother differentiation than rectified linear unit.

$$ELU(\hat{y}) = \begin{cases} \hat{y}_{MHS}, & x > 0 \\ \alpha\left(e^{\hat{y}_{MHS}} - 1\right), & x \leq 0 \end{cases} \quad (1)$$

where α is a negative multiplication factor while $\hat{y}_{MHS}$ represents the result obtained from the multi-head shortcut (MHS).

The positional and contextual attention paradigm gives consecutive encoders temporal attention to encoder output. The encoder outputs a two-dimensional tensor matching input embedding dimensions. Multi-head self-attention, addition, and normalization follow in the decoder module. The first encoder outputs keys and values for the second multi-head attention. The decoder uses the encoder's contextual embedding. Before linear layers, the decoder output flattens. Self-attention and contextual embedding improve decoder understanding and output.

Transformer and BiLSTM architectures are combined with contextual embedding and attention processes in our UPCA technique to estimate RUL, shown in Fig. 1.

### 3.1 Spectral attention (using discrete cosine transform attention)

The encoder of our model utilizes discrete cosine transform attention (DCTA) for its attention

mechanism. DCTA employs the discrete cosine transform (DCT) for converting the signal from the temporal domain to the spectral domain, akin to the Fourier transform. The discrete cosine transform (DCT) utilizes real cosine kernels instead of complex exponential kernels, leading to a real spectrum and adequate energy compaction. A one-dimensional discrete cosine transform (DCT) manages the embedding dimension, denoted as $D_{emb}$. Then, we employ a second one-dimensional DCT, referred to as $D_{seq}$, to manage the sequence dimension.

$$\hat{y} = D_{seq}\left(D_{emb}(q)\right) \tag{2}$$

where $q$ represents query and $\hat{y}$ denoted as DCTA output.

The attention mechanism based on discrete cosine transform (DCT) does not possess any learnable parameters, unlike the mechanisms based on fast Fourier transform (FFT). Reducing parameters improves models' learning capability and practicality when dealing with limited datasets such as C-MAPSS. Additionally, it enhances the efficiency of DCT computation algorithms.

### 3.2 Uninhibited positional and contextual attention (UPCA)

The incorporation noise, sensor data, and positional information in the UPCA model dramatically improves the Transformer's ability to recognize patterns in long-term sequences. Fig. 2 depicts the temporal attention mechanism of the model, which is designed to focus on ordered sequences. Like the discrete cosine transform attention (DCTA), the spectral attention mechanisms incorporate positional encoding into the attention model rather than relying on contextual embedding. This approach maintains the sequential order of the input sequence while considering the surrounding context. One notable characteristic of the UPCA model is its ability to compute temporal attention once and reuse it in subsequent encoders, thereby reducing the computational burden. The two-dimensional temporal attention structure preserves the input and output dimensions, ensuring that time series compression is avoided and that information and order are maintained throughout propagation.

The model utilizes a feed-forward network to produce contextual embeddings by utilizing sliding windows of input sensor data and applying temporal attention. Contextual embedding is a technique that increases the dimensionality of sensors in order to capture complex patterns. Spectral attention

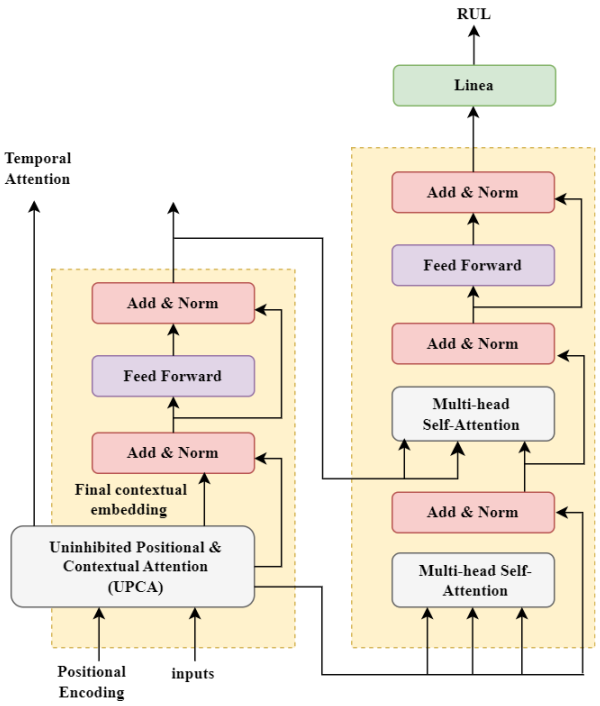


Figure. 1 Proposed frameworks (spectral based transformer)

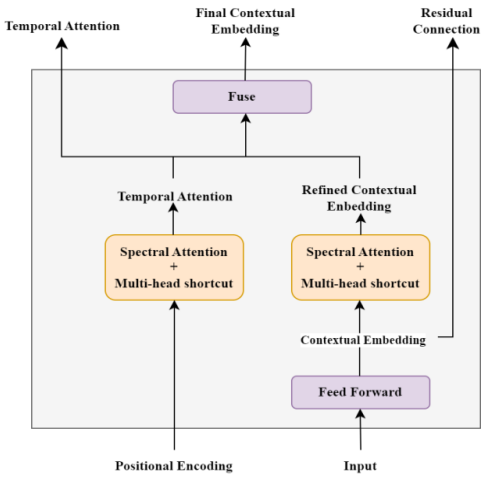


Figure. 2 3.2 Uninhibited positional and contextual attention (UPCA) mechanism

techniques, such as discrete cosine transform attention (DCTA), improve contextual embedding and incorporate a multi-head shortcut mechanism. The merging of enhanced contextual embedding and temporal attention occurs through concatenation along the temporal dimension. Fused the refined contextual embedding with temporal attention by concatenation in temporal dimension:

$$Concat\begin{pmatrix} RefinedContextualEmbedding, \\ TemporalAttention \end{pmatrix} w^F \tag{3}$$

where $w^F$ fusion's weight matrix.

### 3.3 Multi-head shortcuts

This study presents a new method to effectively handle feature diversity and temporal collapse challenges in time series data. Our motivation stems from the desire to utilize an augmented shortcut technique that integrates a multi-head shortcut with a conventional residual connection. The aim is to improve the accuracy of predicting the remaining useful life (RUL) by including various features and maintaining important temporal patterns.

Our approach addresses the problem of low dimensionality in raw sensor data by utilizing a multi-head shortcut with a linear projection layer. This process reduces the loss of information and aids in the identification of subtle patterns that may otherwise go unnoticed. The multi-head shortcut consists of three main components: spectral attention using discrete cosine transformation attention (DCTA), an adjusted residual connection, and a linear layer for feature projection. The integration of these components is achieved through the summation of their outputs. The integration of these components is achieved through the summation of their outputs. The multi-head shortcut ($\hat{y}_{MHS}$) is

$$\hat{y}_{MHS} = \hat{y} + x + w^S \tag{4}$$

where spectral attention's output, input and weight matrix are denoted by $\hat{y}$ , $x$, and $w^S$ respectively.

Our approach enhances remaining useful life (RUL) prediction accuracy by incorporating the multi-head shortcut mechanism and emphasizing the diversity of features. This highlights the significance of including a diverse range of attributes and preserving meaningful patterns over time in the dataset, thereby enhancing the overall accuracy of predictions.

### 3.4 Bidirectional transformer

Bidirectional recurrent neural networks (RNNs) have demonstrated efficacy in sequence modeling through their ability to process input sequences in both forward and reverse directions. This approach improves data understanding by integrating information from both the original and reversed sequences. Two distinct Transformers are implemented, one for forward sequences and another for reversed sequences. The outcomes are aggregated and forwarded through a fully connected layer that utilizes exponential linear unit (ELU) activation and batch normalization. The given representation

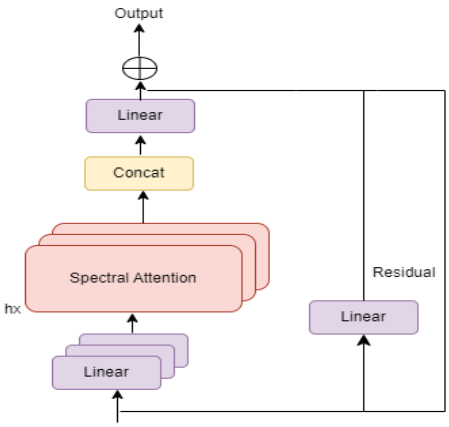


Figure. 3 Multi-head shortcut (MHS)

undergoes a linear layer transformation, resulting in an equivalent sliding window length output. This output estimates the remaining useful life (RUL) for each interval. Fig. 4 illustrates the architecture of the model.

In brief, our methodology employs bidirectional Transformers to capture information from both ends of the input sequence effectively. The aggregated outputs are subjected to additional processing to calculate the remaining useful life (RUL) for specific time intervals.

## 4. Experimental setups

### 4.1 Dataset description and preparation

The C-MAPSS dataset [19] is frequently used in prognostic research and provides valuable insights for developing predictive algorithms. This study entails conducting run-to-failure tests on aircraft engines with different initial wear levels. Data from 21 sensors were collected throughout each cycle until engine failure. The dataset's comprehensive sensor information renders it suitable for algorithm development and evaluation. Table 1 summarizes the four sub-datasets, namely FD001, FD002, FD003, and FD004. These sub-datasets offer unique information that enhances the understanding of the overall dataset, shown in Table 1.

### 4.2 Implementation details

In this study, an embedding dimension of 48 was selected, slightly surpassing the input sensor count of less than 20. The increased input dimensionality enhances the capacity to store and transmit information. The discrete cosine transform algorithm (DCTA) utilizes a query size of 128 and employs a multi-head attention mechanism with two heads to preserve the model size. The encoder and decoder

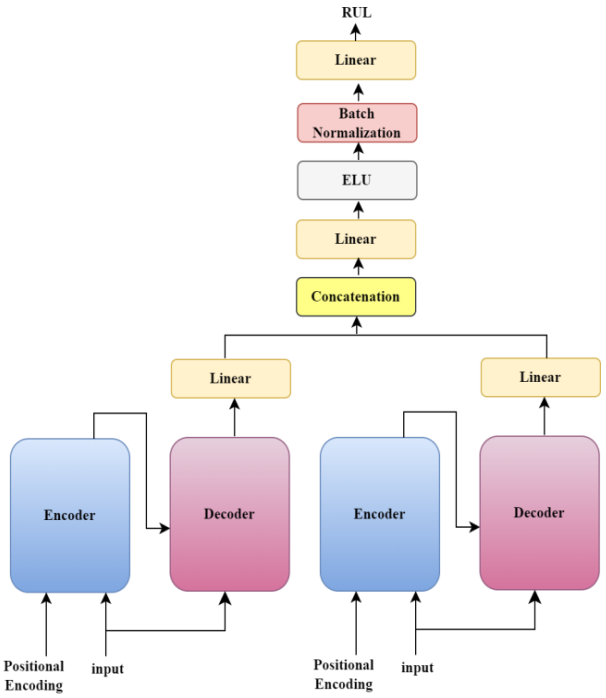


Figure. 4 Transformer with bidirectional input and concatenation

Table 1. Dataset description

| Sub-Datasets | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|
| Train Size | 100 | 260 | 100 | 249 |
| Test Size | 100 | 259 | 100 | 248 |
| Length | 40 | 60 | 40 | 60 |
| Operational | 1 | 6 | 1 | 6 |
| Fault | 1 | 1 | 2 | 2 |
| Training Size | 20,631 | 53,759 | 24,720 | 61,249 |
| Test Size | 100 | 259 | 100 | 248 |

consist of two layers, including an embedded input with a dropout rate of 0.2. The model undergoes 100 training epochs using an Adam optimizer with a learning rate 1e-5 while following Transformer norms. The batch size is 64. The proposed Hybrid model is a more efficient variant of the Transformer architecture, where the encoder component utilizes only one-third of the traditional self-attention mechanism.

Implementation details cover embedding dimension, DCTA query size, multi-head attention headcount, encoder and decoder layers, dropout rate, training optimizer, learning rate, and batch size, all aligned with transformer criteria.

### 4.3 Evaluation

A piecewise linear function generates the ground-truth RUL for a time series. This function uses linear interpolation to predict the remaining cycles until failure ($r_{Linear}$) from the current cycle based on $r_{max}$.

$$r_{Linear} = r_{max} - r_{current} \tag{5}$$

where $r_{current}$ denotes the current cycles.
The piecewise RUL is

$$r = min\,(r_{Linear}, r_{Threshold}) \tag{6}$$

where $r_{Threshold}$ is set to 120.
There are two commonly used evaluation metrics for RUL are RMSE and Score.
Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{r}_i - r_i)^2} \tag{7}$$

$$Score = \begin{cases} \sum_{i=1}^{N}\left(e^{\frac{-\hat{r}_i - r_i}{13}} - 1\right), if \hat{r}_i < r_i \\ \sum_{i=1}^{N}\left(e^{\frac{-\hat{r}_i - r_i}{10}} - 1\right), if \hat{r}_i \geq r_i \end{cases} \tag{8}$$

where $N$ denotes the quantity of samples, $i$ denotes a specific sequence, $\hat{r}_i$ and $r_i$ represent the predicted RUL and the actual RUL, respectively.

## 5. Experimental result analysis

### 5.1 Result analysis of the spectral attention

Table 2 presents a comparison between the "Embedding > Temporal" and "Temporal > Embedding" transformations in the context of Spectral Attention analysis applied to the C-MAPSS dataset. The embedding method consistently produces lower root mean square errors (RMSEs) and scores than the temporal method. For example, in the FD001 dataset, the RMSE for the embedding method is 12.01, with a score of 216, while the temporal method has an RMSE of 12.91 and a score of 233. This trend is also observed in FD002, FD003, and FD004. The results of this study emphasize the superiority of the "Embedding > Temporal" approach in improving the performance of models for predicting Remaining Useful Life (RUL). This highlights the significance of the order of transformations in optimizing the processing of features.

Table 2. Spectral attention

| Datasets | | Embedding > Temporal | Temporal > Embedding |
|---|---|---|---|
| FD001 | RMSE | **12.01** | 12.91 |
| | Score | **216** | 233 |
| FD002 | RMSE | **12.31** | 13.21 |
| | Score | **831** | 3387 |
| FD003 | RMSE | **11.69** | 12.83 |
| | Score | **226** | 253 |
| FD004 | RMSE | **12.78** | 14.34 |
| | Score | **1121** | 1693 |

### 5.2 Ablation study

As presented in Table 3, the research outcomes provide significant insights derived from experimental findings.

The substitution of the discrete cosine transform attention (DCTA) mechanism with conventional self-attention decreased the model's performance. The modification led to a 9% increase in the root mean square error (RMSE) and a 19% increase in the score.

The omission of the uninhibited positional and contextual attention (UPCA) and the fusion model resulted in a significant 60% improvement in score and a 12% decrease in RMSE. The UPCA algorithm can efficiently encode time-series data while effectively reducing the impact of signal embedding noise. Nonparametric positional embedding offers computational benefits in comparison to self-attention.

The system's performance deteriorates when a separate input embedding is used for decoding instead of sharing weights with the encoder. The root mean square error (RMSE) exhibited a 9% increase, whereas the score demonstrated a 72% improvement. The inclusion of input embedding in the initial encoder resulted in improved performance and training outcomes.

Implementing the multi-head shortcut has significantly enhanced the model's performance. The removal of this shortcut resulted in an 11% increase in the root mean square error (RMSE) and a 71% decrease in the score. The multi-head shortcut is a helpful technique that effectively mitigates the problem of feature collapse in deep Transformers. It also enhances the diversity of features in shallow Transformers. This could potentially alleviate the need for extensive manual feature engineering.

The removal of the bidirectional ensemble significantly impacted the score, resulting in a score of 131% after its removal. The mitigation of variation and bias in Bidirectional Transformers is achieved through an ensemble approach, wherein predictions from two lightweight Transformers are averaged.

Including additional encoder and decoder layers did not yield a substantial improvement in the model's performance. Including a more intricate model resulted in a slight enhancement in performance within the C-MAPSS dataset while acknowledging that the existing model's capability was already deemed satisfactory. The absence of any observed performance decline highlights the model's resilience.

These findings emphasize the importance of carefully choosing suitable attention mechanisms in discrete transform and cosine attention (DTCA). The study highlights the importance of UPCA in encoding time-series data, the effectiveness of multi-head shortcuts and bidirectional ensembles, and the model's ability to predict remaining useful life (RUL) accurately.

The Table 3 evaluates component contributions in the proposed CAMPSS dataset model. The "Proposed Model" establishes a baseline using RMSE and score. Substituting traditional Self Attention elevates both RMSE and score. The investigation involves UPCA removal, novel embedding, multi-head shortcut and bidirectional ensemble removal, and additional layers. Relative to the proposed model, percentage change quantifies the effects of each change on predictive accuracy and performance. The results provide crucial insights into each component's specific improvements to the model's outcomes.

### 5.3 Comparison with other State-of-art methods

This study compares the spectral-based transformer (SPT) model's remaining useful life (RUL) prediction with the results presented in the Table 4 and 5. We evaluate our methodology by comparing it to state-of-the-art methods on the complex C-MAPSS dataset, which simulates diverse operating conditions. Our research findings demonstrate that the SPT model excels in hard sub-datasets such as FD002 and FD004. The DCTA mechanism in the novel SPT model significantly improves its performance. The FD002 and FD004 sub-datasets present significant challenges. The SPT model demonstrates a 14% reduction in Root Mean Square Error (RMSE) and a 10% improvement in the complex FD002 dataset score. The SPT model achieves a 15% reduction in root mean square error (RMSE) and a 24% reduction in score on the challenging FD004 dataset. We conduct a comprehensive benchmark by comparing the SPT model with various state-of-the-art methods, each renowned for its distinct contributions.

Table 3. Contribution of each component and compare the results with respect of error (RMSE and Score) in proposed model

| | | Dataset | | | | | |
|---|---|---|---|---|---|---|---|
| | | **FD001** | **FD002** | **FD003** | **FD004** | **Mean** | **% Change** |
| **Proposed Model** | **RMSE** | 12.01 | 12.31 | 11.69 | 12.78 | 12.2 | 0 % |
| | **Score** | 216 | 831 | 227 | 1124 | 599.5 | 0% |
| **Using traditional Self Attention** | **RMSE** | 13.09 | 13.05 | 12.48 | 14.13 | 13.19 | 8% |
| | **Score** | 287 | 1141 | 241 | 1181 | 712.5 | 19% |
| **Removed ICPA** | **RMSE** | 13.03 | 14.04 | 12.69 | 15.12 | 13.72 | 12% |
| | **Score** | 277 | 2019 | 229 | 1321 | 961.5 | 60% |
| **New Embedding** | **RMSE** | 12.81 | 14.12 | 12.42 | 13.93 | 13.32 | 9% |
| | **Score** | 235 | 2438 | 228 | 1214 | 1028.75 | 72% |
| **Removed multi-head shortcut** | **RMSE** | 12.89 | 13.57 | 13.21 | 14.37 | 13.51 | 11% |
| | **Score** | 281 | 2219 | 251 | 1351 | 1025.5 | 71% |
| **Removed Bidirectional** | **RMSE** | 12.3 | 13.27 | 12.7 | 14.76 | 13.26 | 9% |
| | **Score** | 232 | 3608 | 234 | 1457 | 1382.75 | 131% |
| **Adding the Layers** | **RMSE** | 13.17 | 13.1 | 13.28 | 14.65 | 13.55 | 11% |
| | **Score** | 264 | 781 | 247 | 1689 | 745.25 | 24% |

Table 4. Performance comparison based on RMSE

| **Sub-Datasets** | **FD001** | **FD002** | **FD003** | **FD004** |
|---|---|---|---|---|
| Embedded Attention [9] | 12.11 | 15.68 | 12.52 | 18.12 |
| Concurrent-semisupervised [19] | 12.19 | 18.79 | 12.92 | 22.44 |
| DAST [20] | **11.43** | 15.25 | **11.32** | 18.36 |
| Trans+TCNN [21] | 12.31 | 15.35 | 12.32 | 18.35 |
| MSTformer [22] | 12.1 | 14.48 | 12.14 | 15.03 |
| IML Model [23] | 12.42 | 14.03 | 13.39 | 15.10 |
| **Our Proposed** | 12.01 | **12.31** | 11.69 | **12.78** |

Table 5. Performance comparison based on score

| **Sub-Datasets** | **FD001** | **FD002** | **FD003** | **FD004** |
|---|---|---|---|---|
| Embedded Attention [9] | 245 | 1126 | 267 | 2051 |
| con-current-semi super [19] | 208 | 2079 | 245 | 2599 |
| DAST [20] | **203** | 925 | **155** | 1491 |
| Trans+TCNN [21] | 252 | 1267 | 296 | 2120 |
| MSTformer [22] | 207 | 1099 | 248 | 1012 |
| IML Model [23] | 226 | 876 | 227 | 970 |
| **Our Proposed** | 216 | **831** | 227 | **1124** |

The effectiveness of the transformer-based dual aspect self-attention (DAST) [20] technique for remaining useful life (RUL) prediction is enhanced by using dual parallel encoders. Our SPT model incorporates DCTA and UPCA techniques to improve temporal information propagation and effectively handle smaller datasets. Trans+TCNN [21] is a hybrid model that combines transformer and temporal convolutional neural network architectures. It aims to capture long-range dependencies and address overfitting issues in transformer models, particularly in more straightforward scenarios. The SPT model effectively addresses the issue by strategically integrating DCTA and UPCA, enhancing robustness and generality. MSTransformer [22] employs deep learning and attention mechanisms for processing sensor data and operational scenarios. The SPT model balances computing efficiency and performance using DCTA and multi-head shortcut (MHS) techniques. The IML [23] Model facilitates the integration of domain knowledge into RUL forecasts. The SPT model utilizes attention mechanisms to effectively forecast outcomes without requiring domain knowledge integration.

The attention mechanisms and architectural intricacy of the SPT model improve its inherent advantages. The SPT model is well-suited for datasets with limited size. The parameter-free DCTA is a method that reduces memory and computing costs while still enabling accurate predictions with a smaller amount of data. In addition to DCTA, our model incorporates UPCA and MHS. This method enhances prediction accuracy by improving feature extraction, temporal information propagation, and feature representation.

The spectral-based transformer (SPT) model significantly improves remaining useful life (RUL) prediction, particularly in challenging scenarios such as FD002 and FD004. The SPT model is a powerful

machinery prognosis tool due to its innovative attention mechanisms, effective management of restricted datasets, and unmatched performance.

## 6. Conclusion

This study presents the spectral-based transformer (SPT) architecture for predicting remaining useful life (RUL) using the C-MAPSS dataset. Our experiments demonstrate that the SPT model performs better than existing deep learning models, particularly on the challenging FD002 and FD004 sub-datasets. The SPT model is known for its strong stability and ability to generalize well across different sub-datasets, even when trained with limited data. This is due to its advanced components, including discrete cosine transform attention (DCTA), uninhibited positional and contextual attention (UPCA), multi-head shortcuts, and bidirectional ensemble. The SPT model incorporates these innovations to effectively capture complex degradation patterns and enhance the remaining useful life (RUL) prediction accuracy. This results in a 14% reduction in root mean square error (RMSE). Significant performance improvements of 10% for FD002 and 24% for FD004 have been observed. Moreover, the root mean square error (RMSE) for FD004 exhibits a decrease of 15%. Our research provides a comprehensive approach to predicting remaining useful life (RUL), which has the potential to improve maintenance decision-making and enhance system performance in the context of industry 4.0. Future research could involve improving the SPT model to accommodate higher sampling frequencies and expanding its usefulness to different time-series regression tasks. The spectral-based transformer model shows promise for predictive maintenance in the Industry 4.0 era.

## Notation list

| *Variable* | *Description* | *Variable* | *Description* |
|---|---|---|---|
| $\hat{y}_{MHS}$ | *Multi-head Shortcut's output* | $\alpha$ | *Negative Multiplication Factor* |
| $ELU$ | *Exponential Linear Unit* | $\hat{y}$ | *DCTA Output* |
| $D_{emb}$ | *Embedding Dimension* | $q$ | *Query* |
| $w^F$ | *Fusion's weight matrix* | $w^S$ | *Spectral Attention Weight Matrix* |
| $r_{Linear}$ | *Remaining Cycle* | $r$ | *Piecewise RUL* |
| $RUL$ | *Remaining Useful Life* | $RMSE$ | *Root-Mean Square Error* |

## Conflicts of interest

The authors declare no conflict of interest.

## Author contributions

The first author conducted investigations, collected datasets, implemented the study, analyzed the results, and prepared the original draft. The second author provided supervision, completed a work review, and performed validation.

## References

[1] Y. Lei, N. Li, S. Gontarz, J. Lin, S. Radkowski, and J. Dybala, "A Model-Based Method for Remaining Useful Life Prediction of Machinery", *IEEE Transactions on Reliability*, Vol. 65, No. 3, pp. 1314–1326, 2016, doi: 10.1109/tr.2016.2570568.

[2] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications", *Mechanical Systems and Signal Processing*, Vol. 42, Nos. 1–2, pp. 314–334, 2014, doi: 10.1016/j.ymssp.2013.06.004.

[3] H. Soni, A. Kansara, and T. Joshi, "Predictive Maintenance of Gas Turbine using Prognosis Approach", *International Research Journal of Engineering and Technology*, Vol. 07, No. 6, pp. 4683–4691, 2020.

[4] X. S. Si, W. Wang, C. H. Hu, and D. H. Zhou, "Remaining useful life estimation – A review on the statistical data driven approaches", *European Journal of Operational Research*, Vol. 213, No. 1, pp. 1–14, 2011, doi: 10.1016/j.ejor.2010.11.018.

[5] Z. Chen, S. Cao, and Z. Mao, "Remaining Useful Life Estimation of Aircraft Engines Using a Modified Similarity and Supporting Vector Machine (SVM) Approach", *Energies*, Vol. 11, No. 1, p. 28, 2017, doi: 10.3390/en11010028.

[6] Z. Li and Q. He, "Prediction of Railcar Remaining Useful Life by Multiple Data Source Fusion", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 4, pp. 2226–2235, 2015, doi: 10.1109/tits.2015.2400424.

[7] V. Mathew, T. Toby, V. Singh, B. M. Rao, and M. G. Kumar, "Prediction of Remaining Useful Lifetime (RUL) of turbofan engine using machine learning", In: *Proc. of 2017 IEEE International Conference on Circuits and Systems (ICCS)*, 2017, doi:

10.1109/iccs1.2017.8326010.

[8] S. Ahsan and T. A. Lemma, "Remaining Useful Life Prediction of Gas Turbine Engine using Autoregressive Model", *MATEC Web of Conferences*, Vol. 131, p. 04014, 2017, doi: 10.1051/matecconf/201713104014.

[9] X. Zhang, Y. Guo, H. Shangguan, R. Li, X. Wu, and A. Wang, "Predicting remaining useful life of a machine based on embedded attention parallel networks", *Mechanical Systems and Signal Processing*, Vol. 192, p. 110221, Jun. 2023, doi: 10.1016/j.ymssp.2023.110221.

[10] A. A. Dulaimi, S. Zabihi, A. Asif, and A. Mohammadi, "A multimodal and hybrid deep neural network model for Remaining Useful Life estimation", *Computers in Industry*, Vol. 108, pp. 186–196, 2019, doi: 10.1016/j.compind.2019.02.004.

[11] A. L. Ellefsen, E. Bjørlykhaug, V. Æsøy, S. Ushakov, and H. Zhang, "Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture", *Reliability Engineering & System Safety*, Vol. 183, pp. 240–251, 2019, doi: 10.1016/j.ress.2018.11.027.

[12] G. S. Babu, P. Zhao, and X. L. Li, "Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life", *Database Systems for Advanced Applications*, pp. 214–228, 2016, doi: 10.1007/978-3-319-32025-0_14.

[13] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long Short-Term Memory Network for Remaining Useful Life estimation", In: *Proc. of 2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2017, doi: 10.1109/icphm.2017.7998311.

[14] X. Li, Q. Ding, and J. Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks", *Reliability Engineering & System Safety*, Vol. 172, pp. 1–11, 2018, doi: 10.1016/j.ress.2017.11.021.

[15] J. Li, Y. Jia, M. Niu, W. Zhu, and F. Meng, "Remaining Useful Life Prediction of Turbofan Engines Using CNN-LSTM-SAM Approach", *IEEE Sensors Journal*, Vol. 23, No. 9, pp. 10241–10251, 2023, doi: 10.1109/jsen.2023.3261874.

[16] X. Wang, Y. Li, Y. Xu, X. Liu, T. Zheng, and B. Zheng, "Remaining Useful Life Prediction for Aero-Engines Using a Time-Enhanced Multi-Head Self-Attention Model", *Aerospace*, Vol. 10, No. 1, p. 80, 2023, doi: 10.3390/aerospace10010080.

[17] H. Tian, L. Yang, and B. Ju, "Spatial correlation and temporal attention-based LSTM for remaining useful life prediction of turbofan engine", *Measurement*, Vol. 214, p. 112816, 2023, doi: 10.1016/j.measurement.2023.112816.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is All You Need", In: *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[19] T. Wang, D. Guo, and X. M. Sun, "Remaining useful life predictions for turbofan engine degradation based on concurrent semi-supervised model", *Neural Computing and Applications*, Vol. 34, No. 7, pp. 5151–5160, 2021, doi: 10.1007/s00521-021-06089-1.

[20] Z. Zhang, W. Song, and Q. Li, "Dual-Aspect Self-Attention Based on Transformer for Remaining Useful Life Prediction", *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, pp. 1–11, 2022, doi: 10.1109/tim.2022.3160561.

[21] H. K. Wang, Y. Cheng, and K. Song, "Remaining Useful Life Estimation of Aircraft Engines Using a Joint Deep Learning Model Based on TCNN and Transformer", *Computational Intelligence and Neuroscience*, Vol. 2021, pp. 1–14, 2021, doi: 10.1155/2021/5185938.

[22] D. Xu, X. Xiao, J. Liu, and S. Sui, "Spatio-temporal degradation modeling and remaining useful life prediction under multiple operating conditions based on attention mechanism and deep learning", *Reliability Engineering & System Safety*, Vol. 229, p. 108886, 2023, doi: 10.1016/j.ress.2022.108886.

[23] S. Zhou, Y. Yao, A. Liu, F. Wang, L. Chen, and R. Xiong, "Multiform Informed Machine Learning Based on Piecewise and Weibull for Engine Remaining Useful Life Prediction", *Sensors*, Vol. 23, No. 12, p. 5669, 2023, doi: 10.3390/s23125669.